EAI Endorsed Transactions
on Scalable Information Systems

Research Article ICST.ORG

Relationship Among the Diameter of the Area of Influence & Refill Usage of Sri Lanka Using Anonymized Call Detail Records

Wijesinghe, W.O.K.I.S.[1,*], Kumarasinghe, C.U.[1]

[1]Department of Computer Science & Engineering, University of Moratuwa, Sri Lanka

Abstract

Abstract—Economic activity and human mobility are two of the three pillars of socioeconomic indicators and understanding how these correlate with each other is important to society as it may be useful in attempting to classify people into socioeconomic levels using call detail records. Refill usage is one of the key attributes that can taken to model socioeconomic levels as the general sense is that people who spend on more refill are considered as people with high purchasing power. This type of research on SES classification by CDR happened for first time in Sri Lanka and using refill features for modeling, also is not seen in any literature to date. This paper describes what the Diameter of Area of Influence (DAI) is and the relationship between the DAI of an individual which is one of many user mobility features that can be extracted from Call Detail Records (CDRs) and the amount the user refills which is the main economic activity that can be derived from CDRs. This paper also describes a methodology to find DAI using CDRs and how to cluster individual users using this distance.

Received on 12 November 2016; accepted on 18 December 2016; published on 18 January 2017

Keywords: Data Mining, Big Data, Call Detail Records, Socioeconomic Levels, User Mobility, Refill Clustering, Diameter of the Area of Influence.

doi:10.4108/eai.18-1-2017.152104

1. Introduction

Classifying users into socioeconomic levels using call detail records is a cost and time saving method to find current social standing of a society. Finding good set of features to model socioeconomic status is key in creating an accurate model. How to filter out noisy individuals and how to find refill clusters in the set of users is also described in order to elaborate on the correlation between these variables.

2. Data

This research was done on data from one of the telecom providers in Sri Lanka. The data was anonymized by hashing the mobile phone numbers where each hashed mobile number is unique. The following records were used for the data analysis,

- Voice Detail Records - anonymised call numbers, time of initiation & location of initiation (longitude, latitude) derived from a list of Base Transceiver Stations (BTS)
- Refill Detail Records - amount refilled, method of refill & time of refill

The period of study is for five months from January 2012 to May 2012. Data includes both prepaid and post paid users. Data pre-processing was needed to get location data by considering caller direction and removing duplicate records.

3. Methodology

A. *Finding the diameter of the area of influence*

The diameter of the area of influence is the geographical area of influence of an individual's location during their daily activities. In this study it is computed as shown below in Figure 1 rather than

only considering the maximum distance (km) between the set of BTSs used to make or receive calls which was used in [3]. All location data considered was during the period under study.

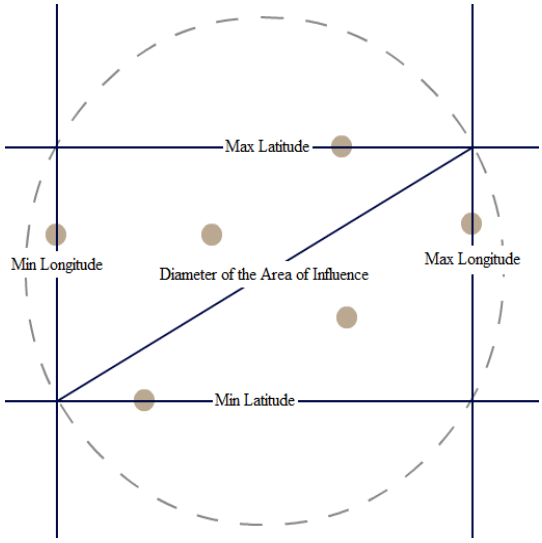

Fig. 1: Diameter of the area of influence calculation: points are the user activity locations which are inside the area of influence dashed circle

The following steps were taken to find the above attribute,

- Find the positions of calls taken by combining voice records and cell id records for each individual
- Find maximum minimum longitude latitude position of each individual during the period of study
- Calculate the diameter of influence using the Haversine Formula for each individual

Below is the equation to derive the haversine distance using earths radius as 6371km,

$$r = 6371km$$
$$lon_{diff} = lon_{max} - lon_{min}$$
$$lat_{diff} = lat_{max} - lat_{min}$$
$$a = (\sin(lat_{diff}/2))^2 + \cos(lat_{max}) * \cos(lat_{min}) + (\sin(lon_{diff}/2))^2$$
$$c = 2 * \arcsin(\min(1, sqrt(a)))$$
$$d = r * c$$

B. Clustering Location Data

In order to cluster users first it is important to filter out users who use their phone less regularly by removing users with average activity of less than 2 per day. For example a user on average who takes or receives less than 2 calls per day is filtered out. This is done in order to get more accurate result as some occasional users may have distorted movement as most of the location data of these users may not be captured. This figure was taken from similar research done [4]

Then a random sample of 100,000 users from these set of frequent users is taken for randomisation and for the ease of computation. In order to cluster the data, first the haversine distance should be normalized. For that the root mean square for the distance is calculated and every distance is divided from the root mean square. Then the normalised distance is centred to the root mean square.

$$rms = \sqrt{(1/n)(x_1^2 + x_2^2 + ... + x_n^2)}$$

Using the k-means algorithm and drawing an elbow graph for k = 1 to 15, it is possible to find the optimal number of clusters in the normalised and centred DAI by finding the elbow point of group sum of squares and by using the that optimal k value as the number of clusters for k-means to label each user by the cluster they belong to.

K-means was used as the clustering algorithms as other algorithms such as DBSCAN required initial parameters which was difficult to calculate and most other research done on similar data preferred k-means. [1][2]

C. Finding Refill Data for each distance cluster

First calculate the total refill amount for each user in Rupees by the summation of all reloads, recharge cards and other methods of refills. Then by selecting the set of users in 100,000 users set taken from haversine distance calculation and joining it with refill data it is possible to find the refill usage of each cluster which can be seen in TABLE II.

D. Clustering Refill Data

It is possible to cluster refill data separately using the same method described in Clustering Location Data section and using the same subset of 100,000 users for the process.

4. Results

Following the methods described in Methodology and by random sampling the below values and figures were calculated. It is also to be noted that all random sample of users taken didn't show any significant variation in results and can be assumed to be true representation of the large set of users.

A. Outcome of DAI clustering

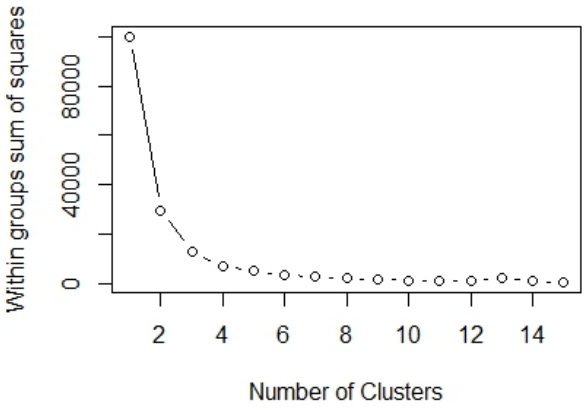

Fig. 2: Elbow curve for identifying best k value for diameter of area of influence

K-means clustering technique applies for clustering the dataset and estimate the optimum number of clusters using elbow curve which shows the group squared sum error (GSSE) versus the number of clusters. The error measure (GSSE) drops monotonically as the number of k clusters increases, but from k=3 the drop flattens significantly as seen by Figure 2.

Max Longitude	Max Latitude	Min Longitude	Min Latitude	DAI (km)	Norm DAI	DAI Cluster
79.91452	6.93727	79.89591	6.93139	2.16	-1.10	2
81.22188	8.602543	79.85437	6.53737	274.99	1.57	3
80.39673	8.328	79.9159	6.790216	179.22	0.63	1
80.7093	7.549812	80.5633	7.26443	35.62	-0.77	2

TABLE I: Final cluster table for diameter of area of influence

The above table displays a sample set of users diameter of their influence, normalised distance and finally the cluster label for each user.

B. Correlation between DAI and refills

The correlation among the diameter of the area of influence value of an individual and the total refill amount spent during the same period was found to be 0.3026. This value was considered after removing low frequency users and post-paid users which was over 80,000 users. If only considering the cluster averages for the correlation between the two attributes then a correlation of 0.9870 can be found from TABLE II.

DAI Cluster Label	Average DAI for each Cluster (km)	Average Refill Amount for each DAI Cluster (Rs.)
1	146.54404	2792.182
2	31.59776	1460.987
3	297.44109	3785.195

TABLE II: Cluster averages for distance and refill

DAI Cluster	Average DAI	Count of Users	Percentage of Users
1	146.54404	35226	35.23%
2	31.59776	48723	48.72%
3	297.44109	16051	16.05%

TABLE III: Cluster distribution for the diameter of the area of influence

From TABLE III, the diameter of the area of influence can be used to classify users as Low, Medium & High using the clusters 2, 1 & 3 respectively.

C. Outcome of refill clustering

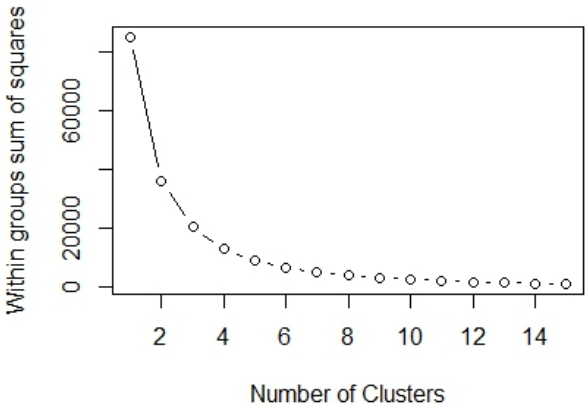

Fig. 3: Elbow curve for identifying best k value for total refill amount

Here the elbow point is not clear as some might say it is 3 where as another might say it is 4.

Refill Cluster Label	Average Total Refill Amount for Refill Cluster (Rs.)	Count of Users	Percentage of Users
R1	869.9941	58625	69.09%
R2	3822.1376	20236	23.85%
R3	22112.0862	690	0.81%
R4	9270.4143	5306	6.25%

TABLE IV: Cluster averages for refill clustering

Clustering using k=4, it is possible to identify from TABLE IV that there are a small number of very high refill users. These may be assumed to be communications, business salesmen etc, which dont represent the society behaviour at large. These then can be removed from our analysis considering them as noise. This then improves the correlation to 0.3384 which is inside the ”good” feature range of 0.3-0.5 described in [3]

5. Conclusion

At individual level it is possible to interpret the results with reasonable confidence (with a correlation of 0.34) that people who have a higher diameter of the area of influence also spend more money for mobile communications and in the case of averaging to the general population, with significant confidence (with a correlation of 0.98) for a developing country like Sri Lanka. This conclusion may be taken forward to use DAI as feature for classifying socioeconomic levels.

References

[1] Richard A Becker, Ramón Cáceres, Karrie Hanson, Ji Meng Loh, Simon Urbanek, Alexander Varshavsky, and Chris Volinsky. Clustering anonymized mobile call detail records to find usage groups. In *Ist Workshop on Pervasive Urban Applications*, 2011.

[2] Richard A Becker, Ramon Caceres, Karrie Hanson, Ji Meng Loh, Simon Urbanek, Alexander Varshavsky, and Chris Volinsky. A tale of one city: Using cellular network data for urban planning. *IEEE Pervasive Computing*, 10(4):18–26, 2011.

[3] Vanessa Frias-Martinez, Jesus Virseda-Jerez, and Enrique Frias-Martinez. On the relation between socio-economic status and physical mobility. *Information Technology for Development*, 18(2):91–106, 2012.

[4] Victor Soto, Vanessa Frias-Martinez, Jesus Virseda, and Enrique Frias-Martinez. Prediction of socioeconomic levels using cell phone records. In *User Modeling, Adaption and Personalization*, pages 377–388. Springer, 2011.